Capstone Project 1 - Milestone Report

Overview

Problem: What is a good indicator for dental health across the globe? Which countries and areas have the largest amount of tooth decay in children under 12 and what is most correlated? The goal of this project is to build models to estimate and predict tooth issues using many demographic factors.

Who is it for: Physicians and organizations who want to understand what kind of demographics and conditions that contribute to dental issues in children. This can then be used to properly allocate funding and focus efforts.

Data used: The data is acquired thru Gapminder and WHO site from the different origins -

- Bad teeth per child (12 yr, WHO)
- GDP/capita (US$, inflation-adjusted, World Bank)
- Government health spending per person (US$, WHO)
- Sugar consumption per person (grams per day, FAO)
- Literacy rate, adult total (% of people ages 15 and above, UNESCO)
- Access to basic drinking water sanitation (% of population with access, WHO)
- Youth crude tobacco use (%, WHO)
- Youth smoking rate (%, WHO)
- Low BMI (% of children more than 2 std below the median BMI, WHO)
- Adolescent birth rate (rate per 1000 women age 15-19, WHO)

Possible Data to add:

- Probability of dying aged 5 to 14 by country (per 1000 children, WHO)
- Number of deaths due to HIV/AIDS by country (Estimates, WHO)

Data Wrangling:
The libraries imported were Pandas and NumPy. So far I have 10 different files that all needed to be cleaned and prepared. Half of them came from a Kaggle dataset and were very similar. These files had information for a large number of countries, each with records over a span of 2 and 20 years, littered with missing data and odd formatting. The other half came from the WHO website. Similar issues needed to be fixed within these files such as dropping missing values, converting dtypes, changing indexes, and then averaged over time. All data after 2004 has been discarded (as the dental data being compared is from 2004). Then correlations in the resulting dataframe are found.

Initial Findings:
There are 10 variables being compared, with 78 countries. There seems to be a statistically significant correlation between bad teeth and adolescent birth rate, access to water sanitation, low bmi in children, sugar consumption, literacy, smokers, and tobacco use.